

Approximating the Average Response Time in Broadcast Scheduling

Nikhil Bansal* Moses Charikar† Sanjeev Khanna‡ Joseph (Seffi) Naor§

Abstract

We consider the problem of approximating the minimum average response time in on-demand data broadcasting systems. The best approximation factors known for this problem involve resource augmentation. We provide the first non-trivial approximation factors in the absence of resource augmentation, achieving an additive $O(\sqrt{n})$ -approximation, where n is the number of distinct pages. Our result can be extended, for any $\epsilon > 0$, to a $(1 + \epsilon)$ -speed, additive $O(1/\epsilon)$ -approximation algorithm. Prior to our work, no non-trivial approximation factor was known for the case of $\epsilon < 1$.

1 Introduction

We consider the problem of minimizing the average response time in *on-demand data broadcasting* systems. In this setting, clients communicate with a powerful server through two independent networks: a network for sending requests to the server (*uplink*), and a “listen only” network from the server to the clients (*downlink*). Typically, the capacity of the downlink is much higher than the capacity of the uplink. Clients send requests to the server for data items that, say, they cannot find locally, and these requests are queued up at the server upon arrival. At each time step, the server chooses a data item among the unsatisfied requests, broadcasts it, and removes the request from the queue. Once a client makes a request, it monitors the downlink until it receives the data item it is waiting for.

Denote the data items by p_1, \dots, p_n . We assume that time is slotted and each data item can be broadcast in a *single* time slot. Let the request sequence be r_1, \dots, r_m . Request r_i is for some particular page p_j and

it arrives at the server at time a_i . Clearly, more than one request can arrive at the same time slot. Denote by b_i the time in which request r_i is satisfied. The *response time* of a request is defined to be the time that elapses from the arrival of the request at the server till the time it is satisfied, i.e., the wait time of request r_i is $b_i - a_i$. We assume that a request cannot be satisfied in the time-slot in which it arrived, thus the minimum response time for any request is at least 1. We want to find a broadcast schedule that minimizes the average response time, defined to be $(\sum_{i=1}^m (b_i - a_i))/m$.

Previous Work and Our results: Our paper focuses on the offline version of the minimum average response time problem, where the request sequence is known in advance to the scheduling algorithm. This problem was shown to be NP-hard by Erlebach and Hall [6]. The algorithmic work on the problem has focused on resource augmentation where the server is given extra speed compared to the optimal algorithm. A k -speed algorithm is one that allows a server to broadcast k pages in each time slot. Kalyanasundaram et al. [9] gave a $\frac{1}{\epsilon}$ -speed, $\frac{1}{1-2\epsilon}$ -approximation algorithm for any fixed ϵ , $0 \leq \epsilon \leq 1/3$. Gandhi et al. [7] have given a $\frac{1}{\alpha}$ -speed, $\frac{1}{1-\alpha}$ -approximation algorithm for any $\alpha \in (0, 1/2]$. Erlebach and Hall [6] gave a 6-speed 1-approximation algorithm for the problem which was improved to a 4-speed 1-approximation algorithm by [7]. Finally, Gandhi et al. [8] give a 3-speed, 1-approximation.

Despite much research on the problem, a non-trivial 1-speed approximation for broadcast scheduling remains an elusive goal. A related question is whether there is an algorithm that achieves a non-trivial approximation guarantee with < 2 -speed.

A common technique for dealing with hard scheduling problems is to consider algorithms that have performance similar to optimum for arbitrarily small speed up factors. Formally, an algorithm is called *fully scalable* if for any arbitrary $\epsilon > 0$, it is a $(1 + \epsilon)$ -speed, $O(1)$ -approximation algorithm¹. Intuitively, a *fully scalable* algorithm guarantees that a system will perform close

*IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598. E-mail: nikhil@us.ibm.com.

†Computer Science Dept., Princeton University, Princeton, NJ 08544. E-mail: moses@cs.princeton.edu. Supported by NSF ITR grant CCR-0205594, DOE award DE-FG02-02ER25540, NSF CAREER award CCR-0237113 and an Alfred P. Sloan Fellowship.

‡Dept. of CIS, University of Pennsylvania, Philadelphia PA 19104. Email: sanjeev@cis.upenn.edu. Supported in part by an Alfred P. Sloan Research Fellowship and by an NSF Career Award CCR-0093117.

§Computer Science Dept., Technion, Haifa 32000, Israel. E-mail: naor@cs.technion.ac.il. Supported in part by the United States-Israel Binational Science Foundation Grant No. 2002-276 and by EU contract IST-1999-14084 (APPOL II).

¹For $0 < \epsilon < 1$, $(1 + \epsilon)$ -speed means that one page can be broadcast in every time slot and an extra page every $1/\epsilon$ time slots.

to optimum unless it is operating close to its full capacity. A detailed discussion on fully scalable algorithms and their implications can be found in [10].

We give here the first non-trivial approximation in the absence of extra speed. We show that an LP-based approach can be used to obtain an additive $O(\sqrt{n})$ -approximation² where n is the number of distinct pages. We also give the first fully scalable algorithm for this problem. In particular, for any $\epsilon > 0$, we give a $(1 + \epsilon)$ -speed, additive $O(1/\epsilon)$ -approximation algorithm.

In the online setting, for the minimum average response time problem, we can show a lower bound of $\Omega(\sqrt{n})$ without speedup and a lower bound of $\Omega(1/\epsilon)$ with a speedup factor of $(1 + \epsilon)$, on the competitive factor of any randomized online algorithm. In [9, Lemma 7], an $\Omega(n)$ lower bound on the competitive ratio of deterministic algorithms is given. Edmonds and Pruhs [4] gave a $(4 + \epsilon)$ -speed, $O(1 + 1/\epsilon)$ -competitive online algorithm. Later, they [5] showed that a natural algorithm, Longest Wait First, is 6-speed, $O(1)$ -competitive. Another measure that has been studied in the literature is minimizing the maximum response time (of a request). For this problem, Bartal and Muthukrishnan [2], gave an $O(1)$ -competitive algorithm.

2 The Algorithm

Our starting point is an integer linear program (ILP) for the broadcast scheduling problem as in [7, 8]. We solve the fractional relaxation of this ILP. The relaxed solution may be viewed as broadcasting pages fractionally at each unit of time such that total fraction of all the pages broadcast in any unit of time does not exceed 1. A request for a page p arriving at a time t is considered satisfied at time t' if t' is the earliest time such that the total amount of page p broadcast during the interval $(t, t']$ is at least 1. The idea of our algorithm is to round this fractional solution by applying a threshold rounding scheme whereby we independently choose a broadcasting threshold $\alpha_p \in [0, 1]$ for each page p . A page p is now scheduled for its first broadcast as soon as α_p units of p are broadcast in the fractional solution. Thereafter, each time an additional unit of fractional broadcast of p is completed, we schedule it for broadcast. Finally, we broadcast the pages in the order in which they are scheduled by this procedure. It is not difficult to show that the expected cost of this schedule is no more than the cost of the fractional relaxation. The problem is that the schedule designed above may not be feasible in that there are multiple pages that are scheduled for broadcast at any given point in time. The

heart of the analysis is to show that with $(1 + \epsilon)$ -speed, the expected backlog at any point in time is $O(1/\epsilon)$, and without speed, it is $O(\sqrt{n})$.

The LP Relaxation: Let $y_{t'}^p = 1$ iff page p is broadcast at time t' , and let $x_{tt'}^p = 1$ iff a request for page p at time t is satisfied at time $t' > t$. Let r_t^p denote the number of requests for page p at time t . Relaxing the integrality constraints on $x_{tt'}$ and y_t gives the following linear program.

$$(2.1) \quad \min \sum_p \sum_t \sum_{t'=t+1}^{T+n} (t' - t) \cdot r_t^p \cdot x_{tt'}^p$$

$$(2.2) \quad \text{subject to } x_{tt'}^p \leq y_{t'}^p, \quad \forall p, t, t' > t$$

$$(2.3) \quad \sum_{t'=t+1}^{T+n} x_{tt'}^p \geq 1, \quad \forall p, t$$

$$(2.4) \quad \sum_p y_{t'}^p = 1, \quad \forall t'$$

$$(2.5) \quad x_{tt'}^p, y_{t'}^p \geq 0 \quad \forall p, t, t'$$

The Rounding Scheme: The algorithm proceeds as follows:

1. Solve the linear program above to obtain an optimal solution of value OPT_{LP} .
2. For each page p , pick $\alpha_p \in [0, 1]$ uniformly and at random (independently for every page).
3. For $i = 0, 1, 2, \dots$, let t_i^p denote the earliest time at which a total of $i + \alpha_p$ fraction of page p has been broadcast in the optimal LP solution.
4. Tentatively schedule a broadcast of page p at times t_0^p, t_1^p, \dots . This schedule is the *tentative schedule*. (Note that it can be infeasible since multiple pages could be broadcast at the same time instant).
5. Assign each tentatively broadcast page to a distinct time slot in a greedy fashion as follows: (a) Any page tentatively broadcast at time t is assigned to a time slot $t' \geq t$. (b) In the final broadcast schedule, pages that are tentatively broadcast at time t precede pages that are tentatively broadcast after time t .

Overview of the Analysis: We break up the analysis of this scheme into two parts. (1) First, we show that the expected wait time of the tentative schedule is at most OPT_{LP} . (Ignoring infeasibility of the tentative schedule). (2) Second, we account for the delay due to the fact that pages that might be actually broadcast later than their tentatively scheduled times. The following property of the rounding scheme will be useful: Consider a time interval $[t_1, t_2]$, where the total fractional broadcast of page p is $\sum_{t=t_1}^{t_2} y_t^p = i + \beta$, where

²Our result is equivalent to a multiplicative $O(\max(1, \sqrt{n}/\text{OPT}))$ approximation.

i is an integer and $\beta \in [0, 1]$. Then, the number of times that page p is tentatively assigned to be broadcast in this interval is either i or $i + 1$. The probability that p is broadcast $i + 1$ times is exactly β .

It is easy to see that for any p, t_1 and t' , if $\sum_{t=t_1+1}^{t'} y_t^p < 1$, then $x_{t_1, t'}^p = y_{t'}^p$, and if $\sum_{t=t_1+1}^{t'} y_t^p \geq 1$, then $x_{t_1, t'}^p = \max(0, 1 - \sum_{t=t_1+1}^{t'-1} y_t^p)$. Otherwise, the solution to the LP can be improved trivially.

The following lemma easily follows from these observations:

LEMMA 2.1. *The expected wait time of the tentative schedule is at most OPT_{LP} .*

Proof. Consider a request for page p that arrives at time t_1 . Let t' be the time this request is satisfied in the tentative schedule. Since α_p is chosen uniformly at random in $(0, 1]$, the probability that $t' \geq z$, is exactly equal to the probability that $\alpha_p > \sum_{t=t_1+1}^{z-1} y_t^p$ which is exactly equal to $\max(1 - \sum_{t=t_1+1}^{z-1} y_t^p, 0)$. Thus the probability that $t' = z$ is exactly $\Pr[t' \geq z] - \Pr[t' \geq z+1] = x_{t_1, z}^p$. It follows that $E[t' - t_1] = \sum_z (z - t_1) x_{t_1, z}^p$, which is exactly the cost of serving this request in the LP solution.

3 Analysis

We view the process of constructing the feasible schedule from the tentative schedule as follows: There is a queue Q , whenever a page p is tentatively scheduled at time t , we add p to Q at time t . At every time step, if Q is non-empty, we broadcast the page at the head of Q . If we are considering the case with a $1 + \epsilon$ speedup, we broadcast 2 pages from the head of Q if t is an integral multiple of $1/\epsilon$ and 1 page otherwise. We will use $Q(t)$ to denote the length of the queue at time t .

Let us consider the response time F_t^p for a request R for page p that arrives at time t . Let t' be the first time after t such that p is scheduled tentatively. Then, F_t^p is at most $t' - t + Q(t')$.

Let $E[F_t^p]$ denote the expected response time for R in the feasible schedule, where the expectation is taken over all the random choices of α_i for all pages i .

Let $\mathcal{E}(p, t, t')$ denote the event that t' is the first time after t when p is scheduled tentatively. Then the above discussion implies that,

$$E[F_t^p] = E[t' - t + Q(t') | \mathcal{E}(p, t, t')]$$

By linearity of expectation, this is equal to $E[t' - t | \mathcal{E}(p, t, t')] + E[Q(t') | \mathcal{E}(p, t, t')]$.

By Lemma 2.1 we know that $E[t' - t | \mathcal{E}(p, t, t')]$ is exactly the LP cost for request R . We will henceforth focus on the second term.

The next lemma shows that conditioning in the second term can be removed without any substantial loss. Intuitively, as t' is the first time after t when p is scheduled tentatively, the choice of t' is only dependent on the choice of α_p . Hence, if we first fix t' and then choose α_p uniformly at random, the only difference is that p may not be tentatively scheduled at t' . We make this precise below.

LEMMA 3.1. $E[Q(t') | \mathcal{E}(p, t, t')] \leq E[Q(t')] + 1$

Proof. Say $p = p_1$. Consider an arbitrary choice of $\alpha_{p_2}, \dots, \alpha_{p_n}$ and fix their values. For any t_1, t_2 , let $A(\alpha, t_1, t_2)$ denote the number of pages that are scheduled tentatively in the interval $(t_1, t_2]$ when $\alpha_{p_1} = \alpha$. Then, for any α and α' such that $0 \leq \alpha, \alpha' \leq 1$, and for any times t_1 and t_2 , it is easy to see that our rounding procedure implies that $|A(\alpha, t_1, t_2) - A(\alpha', t_1, t_2)| \leq 1$.

Thus the value of α_{p_1} can only have limited effect on the number of pages that are scheduled tentatively. We now show that this affects the queue lengths, $Q(t)$, in a limited way too. Formally, we now show that if arrival sequences do not differ (in the sense above) by too much, then queue lengths do not differ either.

Let $A = (a(1), a(2), \dots)$ and $D = (d(1), d(2), \dots)$ be two sequences of non-negative integers. Viewing A as an arrival sequence and D as a departure sequence, A and D determine a queueing system $S(A, D, t)$ as follows: $S(0) = 0$ and $S(t) = \max(0, S(t-1) + a(t) - d(t))$. For our purposes, $a(t)$ will be the number of pages tentatively scheduled at time t and $d(t)$ will denote the number of pages that can be broadcast from the queue Q at time t . In particular, $d(t)$ will be 1 at all times if $\epsilon = 0$. Otherwise if $\epsilon > 0$, then $d(t) = 2$ at times that are integral multiples of $1/\epsilon$ and has the value 1 otherwise.

FACT 3.1. *Let A and \tilde{A} be two sequences such that for any $0 < t_1 < t_2$, $|\sum_{i=t_1+1}^{t_2} (a(i) - \tilde{a}(i))| \leq B$. Let $S = S(A, D, t)$ and $\tilde{S} = S(\tilde{A}, D, t)$, then $|S(t) - \tilde{S}(t)| \leq B$.*

Proof. (of Fact) Let t be the first time where $|S(t) - \tilde{S}(t)| \geq B + 1$. Without loss of generality suppose that $S(t) > \tilde{S}(t)$. Let t_0 be the last time before time t when $S(t) = \tilde{S}(t_0)$. Since then S has had exactly $S(t) + \sum_{x=t_0+1}^t d(x)$ arrivals, so \tilde{S} must also have at least $S(t) + \sum_{x=t_0+1}^t d(x) - B$ arrivals during the interval $(t_0, t]$. So, $\tilde{S}(t)$ must be at least $S(t) - B$, but this gives a contradiction.

Recall that in our setting, we have that $|A(\alpha, t, t') - A(\alpha', t, t')| \leq 1$ for any choice α and α' for α_{p_1} . Thus, setting $B = 1$ in the fact above, we obtain that for every fixed choice of $\alpha_{p_2}, \dots, \alpha_{p_n}$ and any choice of

α and α' for α_{p_1} , the corresponding queue lengths at any time t' differ by at most 1. Taking expectation over all $\alpha_{p_2}, \dots, \alpha_{p_n}$, we have that $|E[Q(t)|(\alpha_{p_1} = \alpha)] - E[Q(t)|(\alpha_{p_1} = \alpha')]| \leq 1$. In particular, choosing α' uniformly at random in $[0, 1]$ and choosing α such that $\mathcal{E}(p, t, t')$ holds, implies that $E[Q(t')|\mathcal{E}(p, t, t')] \leq E[Q(t)] + 1$.

Thus, we will only focus on bounding the expected value of $E[Q(t)]$. In particular, we do not worry about the conditioning on the event $\mathcal{E}(p, t, t')$.

THEOREM 3.1. *In the setting with $(1 + \epsilon)$ -speed, at any time $t = 1, 2, \dots, T + n$, the expected queue size, $E[Q(t)] = O(1/\epsilon)$*

Proof. Fix a $k > \frac{2}{\epsilon}$. We bound the probability that $Q(t) \geq 2k(1 + \epsilon)$. Let us consider the most recent time $t_e < t$ when the queue was idle (i.e. it did not have any page to broadcast).

We divide the time into blocks of intervals of length k . Let I_j denote the time interval $(t - jk, t - (j - 1)k]$, for $j = 1, \dots, \lceil t/k \rceil$. Let η_j denote the event that $t_e \in I_j$. Let $H(p, t_1, t_2)$ denote the number of number of times that p is scheduled tentatively in the interval $(t_1, t_2]$, and let $H(t_1, t_2) = \sum_p H(p, t_1, t_2)$.

Clearly, if η_j holds, then at the end of time $t - (j - 1)k$, the queue length, $Q(t - (j - 1)k)$, can be at most $H(t - jk, t - (j - 1)k)$. Moreover, as the queue always transmits at full capacity after time $t - (j - 1)k$ until time t , it must be the case that at least $(1 + \epsilon)(j - 1)k$ pages are transmit during the interval $t - (j - 1)k$ and t . Since there are $Q(t)$ untransmitted pages at time t , it must be that case that at least $Q(t) + (1 + \epsilon)(j - 1)k - H(t - jk, t - (j - 1)k)$ pages were tentative scheduled during $(t - (j - 1)k, t]$. Hence,

$$H(t - (j - 1)k, t) \geq Q(t) - H(t - jk, t - (j - 1)k) + (1 + \epsilon)(j - 1)k$$

As $H(t - jk, t) = H(t - jk, t - (j - 1)k) + H(t - (j - 1)k, t)$, we have that

$$H(t - jk, t) \geq Q(t) + (1 + \epsilon)(j - 1)k$$

Thus we have that,

$$(3.6) \quad \begin{aligned} Pr[Q(t) \geq 2k(1 + \epsilon)|\eta_j] \\ \leq Pr[H(t - jk, t) - (1 + \epsilon)(j - 1)k \geq 2k(1 + \epsilon)] \end{aligned}$$

We now focus on bounding the right hand side of equation 3.6. Let $s^p(t_1, t_2)$ denote $\sum_{t=t_1+1}^{t_2} y^p(t)$. By our rounding procedure, since each α_p is chosen uniformly in $(0, 1]$, it follows that $H^p(t_1, t_2)$ is $\lceil s^p(t_1, t_2) \rceil$

with probability $\text{frac}(s^p(t_1, t_2))$ and is $\lfloor s^p(t_1, t_2) \rfloor$ otherwise. Hence, $\hat{h}(p, t_1, t_2) = H^p(t_1, t_2) - \lfloor s^p(t_1, t_2) \rfloor$ is a Bernoulli random variable with mean $\text{frac}(s^p(t_1, t_2))$. Thus, we can simplify the right hand side of equation 3.6 as follows.

$$\begin{aligned} & Pr[H(t - jk, t) - (1 + \epsilon)(j - 1)k \geq 2k(1 + \epsilon)] \\ &= Pr[\sum_p \hat{h}(p, t - jk, t) + \sum_p \lfloor s^p(t - jk, t) \rfloor \\ &\quad - (1 + \epsilon)(j - 1)k \geq 2k(1 + \epsilon)] \\ &= Pr[\sum_p \hat{h}(p, t - jk, t) - E[\sum_p \hat{h}(p, t - jk, t)] \\ &\quad + \sum_p s^p(t - jk, t) - (1 + \epsilon)(j - 1)k \geq 2k(1 + \epsilon)] \\ &\leq Pr[\sum_p \hat{h}(p, t - jk, t) - E[\sum_p \hat{h}(p, t - jk, t)] \\ &\quad \geq k(1 + \epsilon) + \epsilon jk] \\ &\leq Pr[\sum_p \hat{h}(p, t - jk, t) - E[\sum_p \hat{h}(p, t - jk, t)] \\ (3.7) \quad &\geq k + \epsilon jk] \end{aligned}$$

The third last step follows as $\sum_p \lfloor s^p(t_1, t_2) \rfloor = \sum_p s^p(t_1, t_2) - \sum_p \text{frac}(s^p(t_1, t_2)) = \sum_p s^p(t_1, t_2) - \sum_p E[\hat{h}(p, t - jk, t)]$ and the second last step follows as $\sum_p s^p(t - jk, t) \leq jk$.

As the events η_j are mutually disjoint events for different j , it follows that

$$Pr[Q(t) \geq 2k(1 + \epsilon)] = \sum_j Pr[Q(t) \geq 2k(1 + \epsilon)|\eta_j] \cdot Pr[\eta_j]$$

and hence that

$$Pr[Q(t) \geq 2k(1 + \epsilon)] \leq \max_j Pr[Q(t) \geq 2k(1 + \epsilon)|\eta_j]$$

Let $p_{j,k}$ denote $Pr[\sum_p (\hat{h}(p, t - jk, t) - E[\hat{h}(p, t - jk, t)]) \geq k + \epsilon jk]$. By equations 3.6 and 3.7 we have that

$$Pr[Q(t) \geq 2k(1 + \epsilon)] \leq \max_j p_{j,k}$$

As a final simplification, for $j \leq 1/\epsilon$, we will ignore the contribution of the term ϵjk in $k + \epsilon jk$ and simply upper bound $p_{j,k}$ by

$$(3.8) \quad Pr[\sum_p (\hat{h}(p, t - jk, t) - E[\hat{h}(p, t - jk, t)]) \geq k]$$

On the other hand, for $j > 1/\epsilon$, we will upper bound $p_{j,k}$ by

$$(3.9) \quad Pr[\sum_p (\hat{h}(p, t - jk, t) - E[\hat{h}(p, t - jk, t)]) \geq \epsilon jk]$$

Since the random variables $\hat{h}(p, t - jk, t)$ are independently distributed for different p , we apply a version of Chernoff bounds as stated in Lemma 3.2 below.³

We are now ready to bound $p_{j,k}$ and hence estimate $E[Q(t)]$.

1. In the case when $j \leq 1/\epsilon$: Clearly for any values of j and k , $E[\sum_p \hat{h}(p, t - jk, t)] \leq E[H(t - jk, t)] \leq jk$ which is at most k/ϵ for $j \leq 1/\epsilon$. Applying lemma 3.2 to equation 3.8, with $\delta = k$ and observing that $\mu \leq jk \leq k/\epsilon$, we get that $p_{j,k} \leq e^{-k\epsilon/4}$.
2. In the case when $j \geq 1/\epsilon$: Applying lemma 3.2 to equation 3.9 with $\delta = \epsilon jk$ and observing that $\mu \leq jk$, we get that

$$p_{j,k} \leq e^{-\epsilon jk \cdot \epsilon/4} < e^{-\epsilon k/4}$$

The last step follows as $j \geq 1/\epsilon$.

Thus for all values of j , we have that,

$$p_{j,k} \leq e^{-\epsilon k/4}$$

and hence that $Pr[Q(t) \geq 2k(1 + \epsilon)] \leq e^{-\epsilon k/4}$ or equivalently $Pr[Q(t) \geq k] \leq e^{-\epsilon k/8(1+\epsilon)}$. Thus,

$$\begin{aligned} E[Q(t)] &= \sum_{k \geq 1} Pr[Q(t) \geq k] \\ &\leq 2/\epsilon + \sum_{k \geq 2/\epsilon} Pr[Q(t) \geq k] \\ &\leq 2/\epsilon + \sum_{k \geq 2/\epsilon} e^{-\epsilon k/8(1+\epsilon)} \\ &= O(1/\epsilon) \end{aligned}$$

This implies the desired result.

LEMMA 3.2. *Let X_1, \dots, X_n be Bernoulli Random variables where p_i denotes the probability that $X_i = 1$. Let $\mu = \sum_i p_i$, then*

$$Pr[\sum_i X_i \leq \mu + \delta] \leq e^{-\delta \min(1/5, \delta/4\mu)}$$

Proof. We use the following additive version of the Chernoff bound (Theorem A.1.10 page 267, [1])

$$Pr[\sum_i X_i \leq \mu + \delta] \leq e^{\delta - \mu \ln(1 + \delta/\mu) - \delta \ln(1 + \delta/\mu)}$$

Let $x = \mu/\delta$, then the right hand side of the term can be written as $e^{-\delta((x+1)\ln(1+1/x)-1)}$. Now, observing that $(x+1)\ln(1+1/x) - 1$ is decreasing in x , and that its value is $3 \ln 5/3 - 1 > 1/5$ at $x = 2$. For $x > 2$, it is at least $(x+1)(1/x - 1/2x^2) - 1 = 1/2x - 1/2x^2 \geq 1/4x$. This gives us the desired bound.

³Since, $\mu = E[\sum_p \hat{h}(p, t - jk, t)]$ can be arbitrarily small in our case, we need to use a special version of the Chernoff bounds, where μ appears in the denominator of the exponent.

3.1 The case when $\epsilon = 0$: We now consider the case when there is no resource augmentation. Setting $\epsilon = 0$ in equation 3.6, we have that

$$Pr[Q(t) \geq 2k|\eta_j] \leq Pr[H(t - jk, t) - (j - 1)k \geq 2k]$$

By equation 3.7 this implies that,

$$\begin{aligned} (3.10) \quad Pr[Q(t) \geq 2k|\eta_j] \\ \leq Pr[\sum_p (\hat{h}(p, t - jk, t) - E[\hat{h}(p, t - jk, t)]) \geq k] \end{aligned}$$

Using the standard Chernoff bound (Theorem A.1.4, page 265 in [1]), we know that if X is the sum of n iid Bernoulli random variables then, $Pr[X - E[X] \geq a] \leq e^{-2a^2/n}$. Since $\sum_p \hat{h}(p, t - jk, t)$ is a sum of at most n Bernoulli random variables we get from equation 3.10 that

$$Pr[Q(t) \geq 2k|\eta_j] \leq e^{-2k^2/n}$$

Finally, as

$$Pr[Q(t) \geq 2k] \leq \max_j Pr[Q(t) \geq 2k|\eta_j]$$

we have that

$$Pr[Q(t) \geq 2k] \leq e^{-2k^2/n}$$

Thus,

$$E[Q(t)] = \sum_{k \geq 1} Pr[Q(t) \geq k] \leq \sum_{k \geq 1} e^{-k^2/2n}$$

Upper bounding $e^{-k^2/2n}$ as $e^{-i^2/2}$ for $k \in [i\sqrt{n} + 1, (i + 1)\sqrt{n}]$, the sum $\sum_{k=1} e^{-k^2/2n}$ is at most

$$\sqrt{n} \cdot \left(\sum_{i=0}^{\infty} e^{-i^2/2} \right) = O(\sqrt{n})$$

Thus we have shown that

THEOREM 3.2. *The randomized algorithm constructs a schedule with $O(\sqrt{n})$ expected additive cost.*

3.2 Derandomization The algorithm above can be derandomized using standard techniques. First, we can assume without loss of generality that α_p are multiples of $1/nT^2$. We claim that this adds at most one to the expected response time of a request. To see this, we can simply round down the values of y_t^p to the closest multiple of $1/nT^2$. The probability that any page p is tentatively scheduled at a different time is at most $nT/nT^2 = 1/T$. In the worst case, we assume that this event simply adds T to the response time of each request. Hence the expected response time of each request goes up by at most 1.

The main idea is that instead of choosing α_p uniformly at random, we choose them from a 4-wise independent family of random variables. As α_p are multiples of $1/nT^2$, there exists such families of size $O(n^4T^8)$. We now show that the expected value of $Q(t)$ at any time t is still $O(\sqrt{n})$ when $\epsilon = 0$ and $O(1/\epsilon)$ when $\epsilon > 0$. The analysis is similar as previously, except that instead of the Chernoff bound, we use the following low independence tail inequalities on l -wise independent random variables, which can be found in [3].

LEMMA 3.3. ([3]) *Let $l \geq 4$ be an even integer. Suppose, X_1, \dots, X_n are l -wise independent random variables taking values in $[0, 1]$. Let $X = X_1 + \dots + X_n$ and $\mu = E[X]$, and let $A > 0$. Then,*

$$(3.11) \quad Pr[|X - \mu| \geq A] \leq 1.004 \left(\frac{nl}{A^2} \right)^{l/2}$$

$$(3.12) \quad Pr[|X - \mu| \geq A] \leq 8 \left(\frac{\mu l + l^2}{A^2} \right)^{l/2}$$

We first consider the case when $\epsilon > 0$. Using the notation in Theorem 3.1, we know that $Pr[Q(t) \geq 2k(1 + \epsilon)] \leq \max_j p_{j,k}$, and by equations 3.8 and 3.9 that

1. For $j \leq 1/\epsilon$, $p_{j,k}$ is at most
$$\leq Pr\left[\sum_p \hat{h}(p, t - jk, t) - E\left[\sum_p \hat{h}(p, t - jk, t)\right] \geq k\right]$$

2. For $j \geq 1/\epsilon$, $p_{j,k}$ is at most,
$$Pr\left[\sum_p \hat{h}(p, t - jk, t) - E\left[\sum_p \hat{h}(p, t - jk, t)\right] \geq \epsilon jk\right]$$

As α_p are 4-wise independent family of random variables, it follows that $\hat{h}(p, t - jk, t)$ are 4-wise independent for different values of p . So, applying equation 3.12 with $l = 4$, $A = k$ and observing that $\mu \leq jk$, we get that

$$p_{j,k} \leq 8 \left(\frac{4jk + 16}{k^2} \right)^2 \leq 8 \left(\frac{20}{\epsilon k} \right)^2 \quad \text{for } j \leq 1/\epsilon$$

For the case when $j > 1/\epsilon$, applying equation 3.12 with $l = 4$ and $A = \epsilon jk$, we have that

$$p_{j,k} \leq 8 \left(\frac{4jk + 16}{\epsilon^2 j^2 k^2} \right)^2 \leq 8 \left(\frac{20}{\epsilon^2 jk} \right)^2 \leq 8 \frac{20^2}{\epsilon^2 k^2}$$

Thus for all j , we have that $p_{j,k} = O(1/k^2 \epsilon^2)$ and hence,

$$Pr[Q(t) > 2k(1 + \epsilon)] = O\left(\frac{1}{\epsilon^2 k^2}\right)$$

Thus,

$$\begin{aligned} E[Q(t)] &= \sum_k Pr[Q(t) > k] \\ &\leq 2/\epsilon + \sum_{k > 2/\epsilon} Pr[Q(t) \geq k] \\ &= 2/\epsilon + \sum_{k > 1/\epsilon} O\left(\frac{1}{\epsilon^2 k^2}\right) \\ &= O(1/\epsilon) \end{aligned}$$

Finally, we consider the case when $\epsilon = 0$. By equation 3.10 we have that,

$$Pr[Q(t) \geq 2k | \eta_j] \leq Pr[\hat{h}(t - jk, t) - E[\hat{h}(t - jk, t)] \geq k]$$

As $\hat{h}(t - jk, t)$ is a sum of at most n 4-wise independent Bernoulli random variables, applying equation 3.11 with $l = 4$ and $A = k$, we get that

$$Pr[Q(t) \geq 2k | \eta_j] \leq 1.004 \left(\frac{4n}{k^2} \right)^2 = O\left(\frac{n^2}{k^4}\right)$$

As, $Pr[Q(t) \geq k] \leq \max_j Pr[Q(t) \geq k | \eta_j]$, we have that $Pr[Q(t) \geq k] = O(n^2/k^4)$. Thus,

$$\begin{aligned} E[Q(t)] &= \sum_k Pr[Q(t) > k] \\ &\leq \sqrt{n} + \sum_{k \geq \sqrt{n}} Pr[Q(t) \geq k] \\ &= \sqrt{n} + \sum_{k \geq \sqrt{n}} O\left(\frac{n^2}{k^4}\right) \\ &= O(\sqrt{n}) \end{aligned}$$

This implies the desired result.

4 Concluding Remarks

In the offline setting, nothing stronger than NP-completeness is known for our problem even in the absence of speed. Perhaps a good starting point is to derive non-trivial lower bounds on the integrality gap of our LP formulation. We can construct an instance where the integrality gap is a small constant, as follows. The request sequence is:

Time 1: one request for page 1 and two requests for page 2.

Time 2: one request for page 2 and one request for page 3.

Time 3: one request for page 1 and one request for page 2.

Time 4: one request for page 3.

An integral solution with cost 14 is 2, 3, 2, 1, 3, starting from time 2. We now show that no other integral solution can have a lower cost. Consider the subproblem consisting of requests at time 3 and 4 only. Any schedule for this must have total response time at least 5. Moreover, any schedule with total response time exactly 5, must transmit either page 1 or page 2 at time 4. Now consider the subproblem consisting of requests at time 1 and 2. It is easy to check that the only solution to this subproblem with total response time 8 is 1, 2, 3, starting from time 2. Note that, as page 3 is transmitted at time 4 in the latter solution, there is no way of combining these two solutions. Thus, any integral solution has total response time at least 14.

An optimal fractional solution is $1/2, 2/3, 1/2, 3, 1/2$, starting from time 2, and where x/y means that one half of page x and one half of page y are broadcast together. The cost of the fractional solution is 13.5, yielding an integrality gap of $14/13.5 \approx 1.027$.

In the online setting, even when randomization is allowed, it is easy to establish lower bounds that (coincidentally) match our offline upper bounds. We outline these lower bounds below.

We first consider the case without speed. At time $t = 0$, the adversary makes r distinct requests, denoted by the set R . At time $t = r/2$, it requests a random subset $R' \subset R$ of size $r/2$. Finally, starting at time r , it requests a new page every unit of time, until time $r + s$. Let S denote this set of requests. An optimal algorithm will first serve the requests in the set $R \setminus R'$ (during the interval $[0, r/2)$), and then requests in R' (during the interval $[r/2, r)$), followed by the requests in the set S (during the interval $[r, r + s)$). The total cost of the optimal algorithm can be upper bounded by $\Theta(r^2) + s$.

To analyze the competitive ratio of a randomized online algorithm, we use Yao's minimax principle and consider the expected cost of any deterministic algorithm A under a request sequence generated by the distribution above. Let X be the set of requests served by A by time $r/2$. Since R' is chosen uniformly at random, the expected size of $(R \setminus X) \cup R'$ is at least $3r/4$. As a result, at time $t = r$, A is expected to have at least $r/4$ unfinished requests from R . The expected cost of A can thus be bounded from below by $(r/4)s$. Choosing $s = r^2$, we can bound the ratio of the expected cost of A to the optimal to be $\Omega(r) = \Omega(\sqrt{n})$.

If the online algorithm has $(1 + \epsilon)$ -speed, then at time $t = r$, expected number of unfinished requests from the set R is $(r/4)(1 - 3\epsilon)$. Choosing $s = \Theta(r/\epsilon)$, we conclude that the online's expected cost is $\Omega(1/\epsilon)$ times the optimal cost.

5 Acknowledgments

We thank Don Coppersmith, Ari Freund, Rajeev Motwani and Kirk Pruhs for several useful discussions.

References

- [1] N. Alon, and J. Spencer, *The Probabilistic Method*. John Wiley & Sons, 2000.
- [2] Y. Bartal and S. Muthukrishnan. *Minimizing maximum response time in scheduling broadcasts*. Proc. of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 558-559, 2000.
- [3] M. Bellare and J. Rompel. *Randomness-Efficient Oblivious Sampling*. Proc. of the 35th Annual IEEE Symposium on Foundations of Computer Science, pp. 276-287, 1994.
- [4] J. Edmonds and K. Pruhs. *Broadcast scheduling: when fairness is fine*. Proc. of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 421-430, 2002.
- [5] J. Edmonds and K. Pruhs. *A maiden analysis of Longest Wait First*. Proc. of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 818-827, 2004.
- [6] T. Erlebach and A. Hall. *NP-hardness of broadcast scheduling and inapproximability of single-source unsplittable min-cost flow*. Proc. 13th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 194-202, 2002.
- [7] R. Gandhi, S. Khuller, Y. Kim and Y-C. Wan. *Approximation algorithms for minimizing response time in broadcast scheduling*. Proc. of the 9th Conference on Integer Programming and Combinatorial Optimization (IPCO), pp. 425-438, 2002.
- [8] R. Gandhi, S. Khuller, S. Parthasarathy and A. Srinivasan. *Dependent Rounding on Bipartite Graphs*. Proc. of 43rd IEEE Conference on Foundations of Computer Science (FOCS), pp. 323-332, 2002.
- [9] B. Kalyanasundaram, K. Pruhs, and M. Velauthapillai, *Scheduling broadcasts in wireless networks*, Proc. of the 8th Annual European Symposium on Algorithms, pp. 290-301, 2000.
- [10] K. Pruhs, J. Sgall, E. Torng, *Online Scheduling*. Handbook of Scheduling: Algorithms, Models, and Performance Analysis, editor Joseph Y-T. Leung, CRC Press, 2004.